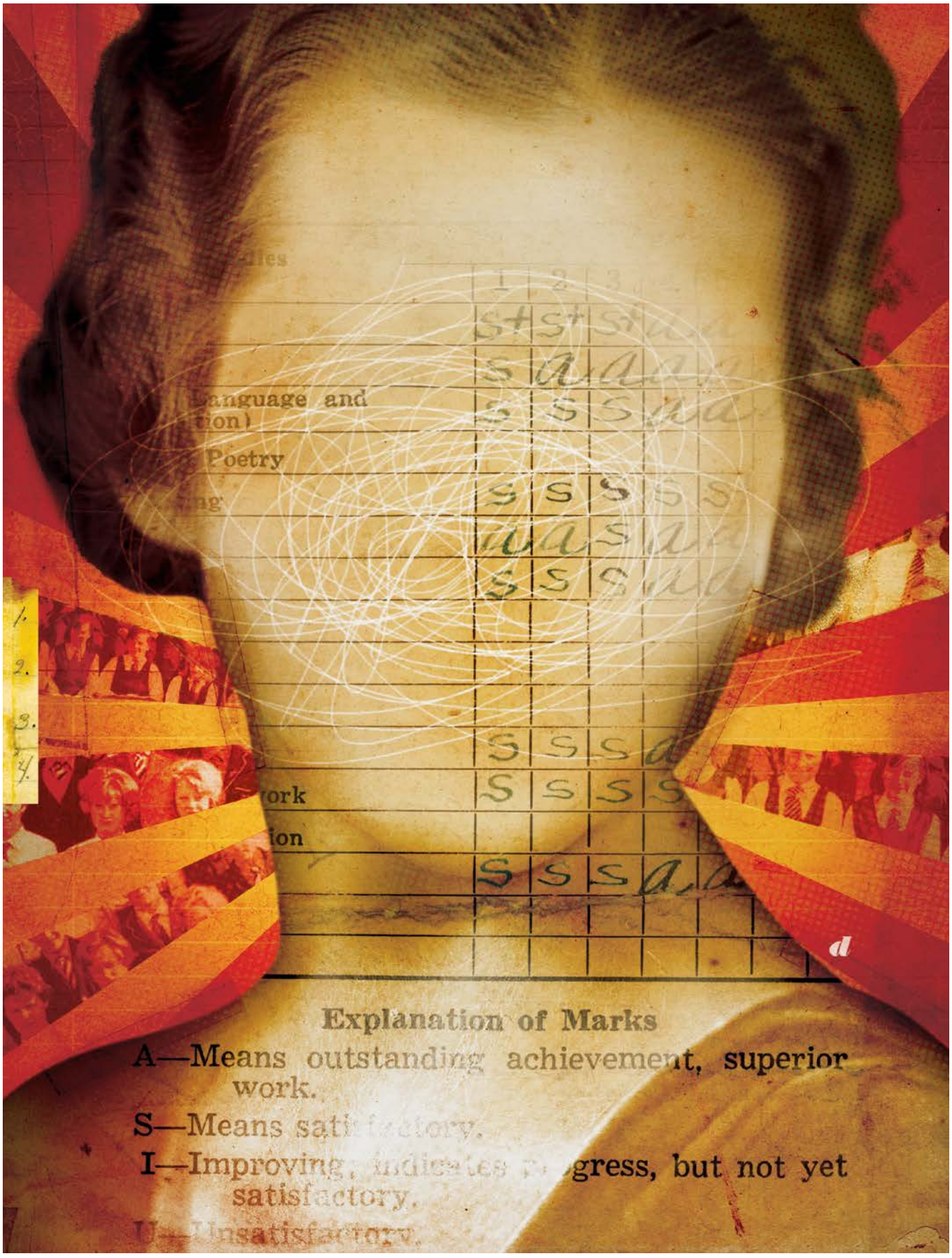


ILLUSTRATION / DAVID VOGIN

Getting Classroom Observations RIGHT

Lessons on How from Four Pioneering Districts

IT IS WIDELY UNDERSTOOD that there are vast differences in the quality of teachers: we've all had really good, really bad, and decidedly mediocre ones. Until recently, teachers were deemed qualified, and were compensated, solely according to academic credentials and years of experience. Classroom performance was not considered. In the last decade, researchers have used student achievement data to quantify teacher performance and thereby measure differences in teacher quality. Among the recent findings is evidence that having a better teacher not only has a substantial impact on students' test scores at the end of the school year, but also increases their chances of attending college and their earnings as adults (see "Great Teaching," *research*, Summer 2012).

In response to these findings, federal policy goals have shifted from ensuring that all teachers have traditional credentials and are fully certified to creating incentives for states to evaluate and retain teachers based on their classroom performance. We contribute to the body of knowledge on teacher evaluation systems by examining the actual design and performance of new teacher-evaluation systems in four school districts that are at the forefront of the effort to evaluate teachers meaningfully.

We find, first, that the ratings assigned teachers by the districts' evaluation systems are sufficiently predictive of a teacher's future performance to be used by administrators for high-stakes decisions. While evaluation systems that make use of student test scores, such as value-added methods, have been the focus of much recent debate, only a small fraction of teachers, just one-fifth in our four study districts, can be evaluated based on gains in their students' test scores. The other four-fifths of teachers, who are responsible for classes not covered by standardized tests, have to be evaluated some other way, including, in our districts, by basing the teacher's evaluation score on classroom observations, achievement test gains for the whole school, performance on nonstandardized tests chosen and administered by each teacher to her own students, and by some form of "team spirit" rating handed out by administrators. In the four districts in our study, classroom observations carry the bulk of the weight, comprising between 50 and 75 percent of the overall evaluation scores for teachers in non-tested grades and subjects.

As a result, most of the action and nearly all the opportunities for improving teacher evaluations lie in the area of

by GROVER J. "RUSS" WHITEHURST, MATTHEW M. CHINGOS, AND KATHARINE M. LINDQUIST

classroom observations rather than in test-score gains. Based on our analysis of system design and practices in our four study districts, we make the following recommendations:

1) Teacher evaluations should include two to three annual classroom observations, with at least one of those observations being conducted by a trained observer from outside the teacher's school.

2) Classroom observations that make meaningful distinctions among teachers should carry at least as much weight as test-score gains in determining a teacher's overall evaluation score when both are available.

3) Most important, districts should adjust teachers' classroom-observation scores for the background characteristics of their students, a factor that can have a substantial and unfair influence on a teacher's evaluation rating. Considerable technical attention has been given to wringing the bias out of value-added scores that arises because student ability is not evenly distributed across classrooms (see "Choosing the Right Growth Measure," *research*, Spring 2014). Similar attention has not been paid to the impact of student background characteristics on classroom-observation scores.

Most of the action and nearly all the opportunities for improving teacher evaluations lie in the area of classroom observations rather than in test-score gains.

Observations vs. Value-Added

The four urban districts we study are scattered across the country. Their enrollments range from about 25,000 to 110,000 students, and the number of schools ranges from roughly 70 to 220. We have from one to three years of individual-level data on students and teachers, provided to us by the districts and drawn from one or more of the years from 2009 to 2012. We begin our analysis by examining the extent to which the overall ratings assigned to teachers by the districts' evaluation systems are predictive of the teacher's ability to raise test scores and the extent to which they are stable from one year to the next. The former analysis can be conducted only for the subset of teachers with value-added ratings, that is, teachers in tested grades and subjects. In contrast, we can examine the stability of overall ratings for all teachers included in the districts' evaluation systems.

We find that the overall evaluation scores in one year are correlated with the same teachers' value-added scores in an adjacent year at levels ranging from 0.33 to 0.38. In other words, teacher-evaluation scores based on a number of components, including teacher- and school-level value-added scores, classroom-observation scores, and other student and administrator ratings, are quite predictive of a teacher's ability to raise student test scores the following (or previous) year. The year-to-year correlation is in keeping with the findings from prior research on value-added measures when used on their own as a teacher-performance metric. The degree of correlation confirms that these systems perform substantially better in predicting future teacher performance than traditional systems based on paper credentials and years of experience. These correlations are also in the range that is typical of systems for evaluating and predicting future performance in other fields of human endeavor, including, for example, those used to make management decisions on player contracts in professional sports.

We calculate the year-to-year stability of the evaluation scores as the correlation between the overall scores of the same teachers in adjacent years. The stability generated by the districts' evaluation systems ranges from a bit more than 0.50 for teachers with value-added scores to about 0.65 when value-added is not a component of the score. Evaluation scores that do not include value-added are more stable because they assign more weight to observation scores, which are more stable over time than value-added scores.

Why are observation scores more stable? The difference may be due, in part, to observations typically being conducted by school administrators who have preconceived ideas about a teacher's effectiveness. If a principal is positively disposed toward a particular teacher because of prior knowledge, the teacher may receive a higher observation score than the teacher would have received if the principal were unfamiliar with her or had a prior negative disposition. If the administrator's impression of individual teachers is relatively sticky from year to year, then it will be less reflective of true teacher performance as observed at a particular point of time. For this reason, maximizing stability may not increase the effectiveness of the evaluation system.

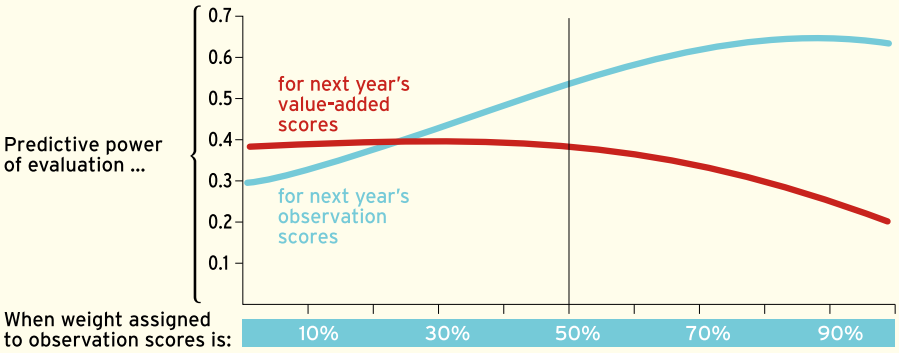
This leaves districts with important decisions to make regarding the tradeoff between the weights they assign to value-added versus observational components for teachers in tested grades and subjects. Our data show that there is a tradeoff between predicting observation scores and predicting value-added scores of teachers in a subsequent year. Figure 1 plots the ability of an overall evaluation score, computed based on a continuum of different weighting schemes, to predict teachers' observation

and value-added scores in the following year. The optimal ratio of weights to maximize predictive power for value-added in the next year is about two to one (value-added to observations), whereas maximizing the ability to predict observations requires putting the vast majority of weight on observations.

We do not believe there is an empirical solution for the ideal weights to assign to observation versus value-added scores. The assignment of those weights depends on the a priori value the district assigns to raising student test scores, the confidence it has in its classroom-observation system as a tool for both evaluation and professional development, and the political and practical realities it faces in negotiating and implementing a teacher-evaluation system.

Finding a Balance (Figure 1)

If less than 50 percent of the weight in an evaluation is assigned to observation scores, significant increases in the ability to predict next year's observation scores can be obtained by increasing the weight assigned to observations with relatively little decrease in the ability to predict next year's value-added scores.



NOTE: This analysis assumes that the remaining weight in the evaluation is placed on value-added scores.

SOURCE: Authors' calculations

At the same time, there are ranges of relative weighting—namely between 50 and 100 percent value-added—where significant increases in the ability to predict observation scores can be obtained by increasing the weight assigned to observations with relatively little decrease in the ability to predict value-added. Consequently, most districts considering only these two measures should assign a weight on observations of at least 50 percent.

Classroom Observation

The structure, frequency, and quality of the classroom observation component are also important. The observation system in place should make meaningful distinctions among teachers. An observation system that provides only two choices, satisfactory and unsatisfactory, for example, will result in similar ratings being given to most teachers. As a result, the observation component will not carry much weight in the overall evaluation score, regardless of how much weight is officially assigned to it. An evaluation system with little variation in observation scores would also make it very difficult for teachers in nontested grades and subjects to obtain very high or low total evaluation scores; they would all tend to end up in the middle of the pack, relative to teachers for whom value-added scores are available.

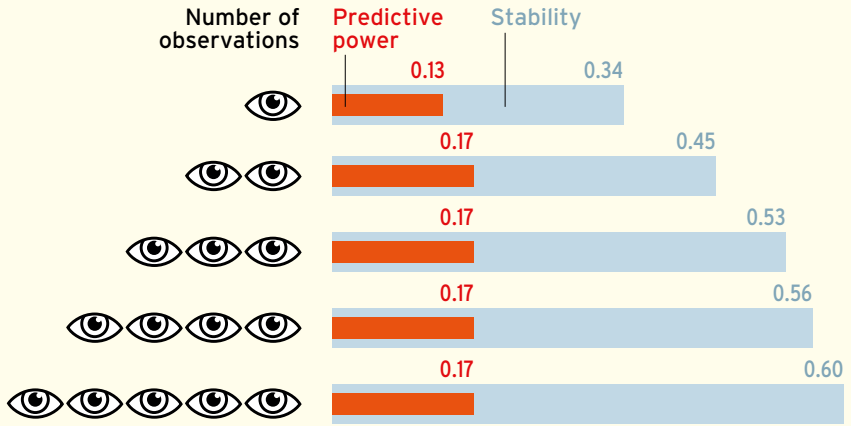
In all of our study districts, the quality of information garnered from classroom observations depends on how many are conducted. Figure 2 shows that moving from one to two observations (independent of who conducts them) increases both the stability of observation scores and their predictive power for value-added scores in the next year. Adding additional observations continues to increase the stability of observation scores but has no further effect on their predictive power for future value-added scores.

In districts that use a mix of building leaders and central administration staff to conduct classroom observations, the quality of the information also depends on who conducts the observations. Observations conducted by in-building administrators, e.g., the principal, are more stable (0.61) than those done by central administration staff (0.49), but observations conducted by evaluators from outside the building have higher predictive power for value-added scores in the next year (0.21) than those done by administrators in the building (0.15). The higher year-to-year stability of observations conducted by the principal or assistant principal compared to out-of-building observers is consistent with our hypothesis that a principal's observation is influenced by both her preexisting opinion about a given teacher and the information that is derived from the classroom observation itself.

Classroom observations are expensive, and, for the majority of teachers, they are the most heavily weighted contributor to their individual evaluation score. Observations also have a critical role to play for principals who intend to be instructional leaders, as they present a primary point of contact between

Optimal Observation (Figure 2)

Moving from one to two observations each year increases both the stability of observation scores and their predictive power for value-added scores in the next year.



NOTE: Predictive power is the correlation between a teacher's observation scores and her value-added the following year; stability is the correlation between a teacher's observation score and her observation score in adjacent years.

SOURCE: Authors' calculations

the school leader and classroom teaching and learning. It is important to balance what are, in part, the competing demands of empowering the school leader to lead, spending no more than necessary in staff time and money to achieve an effective observation system, and ensuring that observation scores are based on what is observed rather than on extraneous knowledge and prior relationships between the observer and the teacher.

Our data suggest that three observations provide about as much value to administrators as five. We recommend that districts conduct two to three annual classroom observations for each teacher, with at least one of those being conducted by a trained observer from outside the teacher's school without substantial prior knowledge of, or conflict of interest with respect to, the teacher being observed. Districts should arrange for an additional classroom observation by another independent observer in cases in which there are substantial and potentially consequential differences between the observation scores generated by the primary observers.

Bias in Observation Scores

A teacher-evaluation system would clearly be intolerable if it identified teachers in the gifted and talented program as superior to other teachers because students in the gifted and talented program got higher scores on end-of-year tests. Value-added metrics mitigate this bias by measuring test-score gains from one school year to the next, rather than absolute scores at the end of the year, and by including statistical controls for characteristics of students and classrooms that are known to be associated with student test scores, such as students' eligibility for free or reduced-price lunch.

But as noted above, classroom observations, not test-score gains, are the major factor in the evaluation scores of most teachers in the districts we examined, ranging from 40 to 75 percent of the total score, depending on the district and whether the teacher is responsible for a classroom in a tested grade and subject. Neither our four districts, nor others of which we are aware, have processes in place to address the possible biases in observation scores that arise from some teachers being assigned a more-able group of students than other teachers.

Imagine a teacher who, through the luck of the draw or administrative decision, gets an above-average share of students who are challenging to teach because they are less well prepared academically, aren't fluent in English, or have behavioral problems. Now think about what a classroom observer is asked to judge when rating a teacher's ability. For example, in a widely used classroom-observation system created by Charlotte Danielson, a rating of "distinguished" on questioning and discussion techniques requires the teacher's questions to consistently provide high cognitive challenge with adequate time for students to respond, and requires that students formulate many questions during discussion. Intuitively, the teacher with a greater share of students who are challenging to teach is going to have a tougher time performing well under this rubric than the teacher in the gifted and talented classroom.

This intuition is borne out in our data: teachers with students with higher incoming achievement levels receive classroom- observation scores that are higher on average than those received by teachers whose incoming students are at lower achievement levels. This finding holds when comparing the observation scores of the same teacher with different classes of students. The latter finding is important because it indicates that the association between student incoming

achievement levels and teacher-observation scores is not due, primarily, to better teachers being assigned better students. Rather, it is consistent with bias in the observation system; when observers see a teacher leading a class with higher-ability students, they judge the teacher to be better than when they see that same teacher leading a class of lower-ability students.

Figure 3 depicts this relationship using data from teachers in tested grades and subjects for whom it is possible to examine the association between the achievement levels of students that teachers are assigned and the teachers' classroom-observation scores. Notice that only about 9 percent of teachers assigned a classroom of students who are "lowest achieving" (in the lowest fifth of academic performance based on their incoming test scores) are identified as top-performing based on classroom observations, whereas the expected outcome would be 20 percent if there were no association between students' incoming ability and a teacher's observation score. In contrast, four times as many teachers (37 percent) whose incoming students are "highest achieving" (in the top fifth of achievement based on incoming test scores) are identified as top performers according to classroom observations.

This represents a serious problem for any teacher-evaluation system that places a heavy emphasis on classroom observations, as nearly all current systems are forced to do because of the lack of measures of student learning in most grades and subjects. Fortunately, there is a straightforward fix to this problem: adjust teacher-observation scores based on student background characteristics, which, unlike prior test scores, are available for all teachers. We implement this adjustment using a regression analysis that calculates each teacher's observation score relative to her predicted score based on the composition of her class, measured as the percentages of students who are white, black, Hispanic, special education, eligible for free or reduced-price lunch, English language learners, and male. These background variables are all associated with entering achievement levels, but we do not adjust for prior test scores directly because doing so is only possible for the minority of teachers in tested grades and subjects.

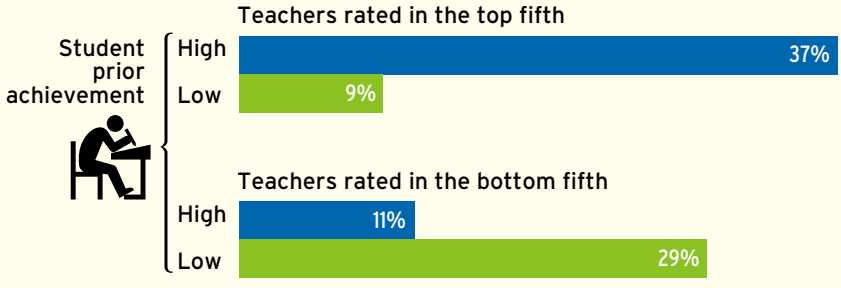
Moving from one to two observations increases both the stability of observation scores and their predictive power for value-added scores.

Such an adjustment for the makeup of the class is already factored in for teachers for whom value-added is calculated, because student gains are adjusted for background characteristics. But the adjustment is only applied to the value-added portion of the evaluation score. For the teachers in nontested grades and subjects, whose overall evaluation score does not include value-added data, there is no adjustment for the makeup of their class in any portion of their evaluation score.

When classroom-observation scores are adjusted for student background characteristics, the pattern of observation scores is much less strongly related to the incoming achievement level of students than is the case when raw classroom-observations scores are used. A statistical association remains between incoming student achievement test scores and teacher ratings based on classroom observations, but it is reduced substantially.

Unfair Advantage (Figure 3)

When rated based on classroom observations, four times as many teachers of students with high prior achievement (37 percent) than teachers of low-achieving students (9 percent) are deemed top performers (that is, rated in the top 20 percent of all teachers). This suggests that observers are apt to assign high ratings to teachers they see leading high-ability classrooms, regardless of the teachers' actual performance.



SOURCE: Authors' calculations

States have an important role to play in helping local districts make these statistical adjustments. In small districts, small numbers of students and teachers will make these kinds of adjustments very imprecise. We estimate the magnitude of this issue by creating simulated small districts from the data on our relatively large districts, and find that the number of observations in the adjustment model can have a large impact on the stability of the resulting evaluation measures.

The solution to this problem is also straightforward. States should conduct the statistical analysis used to make adjustments using data from the entire state, or subgroups of demographically similar districts, and provide the information necessary to calculate adjusted observation scores back to the individual districts. The small number of states that already have evaluation systems in place address this issue by calculating value-added scores centrally and providing them to local school districts. This should remain the norm and be expanded to include observation scores, given the important role they play in the evaluations of all teachers.

Districts should adjust teachers' classroom-observation scores for the background characteristics of their students, a factor that can have a substantial and unfair influence on a teacher's evaluation rating.

Conclusions

A new generation of teacher-evaluation systems seeks to make performance measurement and feedback more rigorous and useful. These systems incorporate multiple sources of information, including such metrics as systematic classroom observations, student and parent surveys, measures of professionalism and commitment to the school community, more differentiated principal ratings, and test-score gains for students in each teacher's classrooms.

Although much of the impetus for new approaches to teacher evaluation comes from policymakers at the state and national levels, the design of any particular teacher-evaluation system in most states falls to individual school districts and charter schools. Because of the immaturity of the knowledge base on the design of teacher-evaluation systems, and the local politics of school management, we are likely to see considerable variability among school districts in how they go about evaluating teachers.

That variability is a double-edged sword. It offers the opportunity to study and learn from natural variation in the design of evaluation systems, but it also threatens to undermine public support for new teacher-evaluation systems to the extent that the natural variation suggests chaos, and is used by opponents of systematic teacher evaluation to highlight the failures of the worst-performing systems. The way out of this conundrum is to accelerate the process by which we learn from the initial round of district experiences and to create leverage points around that learning that will lift up the weakest evaluation systems.

Our examination of the design and performance of the teacher-evaluation systems in four districts provides reasons for optimism that new, meaningful evaluation systems can be designed and implemented by individual districts. At the same time, we find that our districts share, to one degree or another, design decisions that limit their systems' performance, and that will probably be seen as mistakes by stakeholders as more experience with the systems accrues. We focus our recommendations on improving the quality of data derived from classroom observations. Our most important recommendation is that districts adjust classroom observation scores for the degree to which the students assigned to a teacher create challenging conditions for the teacher. Put simply, the current observation systems are patently unfair to teachers who are assigned less-able and -prepared students. The result is an unintended but strong incentive for good teachers to avoid teaching low-performing students and to avoid teaching in low-performing schools.

A prime motive behind the move toward meaningful teacher evaluation is to assure greater equity in students' access to good teachers. A teacher-evaluation system design that inadvertently pushes in the opposite direction is clearly undesirable. We have demonstrated that these design errors can be corrected with tools in hand.

Grover J. "Russ" Whitehurst is director of the Brown Center on Education Policy at the Brookings Institution, where Matthew M. Chingos is a senior fellow, and Katharine M. Lindquist is a research analyst.